Contents lists available at SciVerse ScienceDirect

Food and Bioproducts Processing

journal homepage: www.elsevier.com/locate/fbp

Utilisation of key descriptors from protein sequence data to aid bioprocess route selection

Christopher J. O'Malley[a], *Gary A. Montague*[a,*], *Elaine B. Martin*[a], *John M. Liddell*[b], *Bo Kara*[b], *Nigel J. Titchener-Hooker*[c]

[a] *School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK*
[b] *Fujifilm Diosynth Biotechnologies (FDB), Belasis Avenue, Billingham, TS23 1LH, UK*
[c] *Department of Biochemical Engineering, University College London, London WC1E 6BT, UK*

A B S T R A C T

The large-scale manufacture of biological products results in the generation of significant quantities of process information that can be used to inform future design decisions. Currently this information is not exploited to its full potential. The challenge is thus to identify and/or develop tools that allow the utilisation of this valuable resource. The main objective of the research reported in this paper was to investigate whether it was possible to utilise information, in particular that extracted from protein sequence data, from previous processes, with the goal of informing process route selection early in development. The approach adopted draws on tools in the areas of data mining and pattern recognition including the techniques of Fisher correlation score and self-organising maps. The methodology developed was applied to two case studies utilising data from the amino acid sequences of 41 proteins previously developed at Avecia Biologics, along with associated information relating to the downstream processing steps used during their large scale manufacture. The results demonstrate that information from previous processes can be used to inform process route selection.

Keywords: Bioprocess design; Protein sequence descriptors; Self-organising feature maps; Kohonen networks; Industrial application; Variable selection; Clustering; Protein sequence data; Process route selection; Data mining

1. Introduction

The development and manufacture of human healthcare products is an exciting and fast moving sector of the biotechnology industry. However, to be successful in this competitive environment, biotech companies need to balance the conflicting business requirements resulting from the need to minimise time to market, while at the same time ensuring product is fit for purpose and maximising production efficiency. Success relies on creating an environment where maximum use is made of available knowledge and expertise. This is particularly important in the area of contract development and manufacture, where time lines are short and product histories are limited.

Over the past 30 years, the biopharmaceutical industry has launched more than 100 molecules, a figure which is anticipated to maintain a growth rate of 15–30% annually (Walsh, 2006). For whole antibodies (IgG), which have been a major growth area of biopharmaceuticals for a number of years, there exist template/platform processes based on protein A capture (Shukla et al., 2007). Consequently the basic process structure is in place for any IgG, with development focussing on the optimisation of the platform for individual cases. However, for therapeutic proteins, as a general category, no platform exists and development of a process for each new protein must be treated on an individual basis. Major fundamental process design decisions that need to be made include consideration of:

* Corresponding author.
E-mail address: Gary.Montague@ncl.ac.uk (G.A. Montague).
Received 7 August 2011; Received in revised form 13 January 2012; Accepted 19 January 2012
0960-3085/$ – see front matter
doi:10.1016/j.fbp.2012.01.005

- Organism (prokaryote, eukaryote)
- Location of product accumulation (intracellular, periplasmic, secreted)
- Strain
- Construct
- Primary separation methods (centrifugation, TFF, direct capture)
- Multi-step chromatographic purification (chromatography media selection, operating conditions, elution strategy, pooling criteria, etc.)
- Drug substance formulation

At production scale, a complex series of options for process route selection exist. The potential to assess experimentally the multiple options available for route selection and development has resulted in the need to generate data that would support process development. One approach to receive significant attention is that of high throughput process development involving extremely scaled down, parallel experimentation (Micheletti and Lye, 2006). An alternative, complementary approach is to make use of the knowledge contained in data from pre-existing processes (Avramenko and Kroslawski, 2006). This can range from relatively small amounts of data from early stage design, such as from shake flask experiments, to more comprehensive, later phase records such as those generated during scale-up studies. Accessing and interpreting the underlying correlation structure of this valuable resource have the potential to enhance process knowledge and understanding and hence contribute to route selection and process development.

The nature of the complex processes used at production scale makes a mechanistic approach to understanding the fundamental relationships of bioprocess data a challenge, with reduced complexity models being the only viable option. A more industrially relevant approach is to adopt a data based style of knowledge acquisition, specifically, through the utilisation of data mining and pattern recognition techniques, for example. Typically, these are less labour-intensive and time-consuming than mechanistic approaches. By adopting such methods, a degree of inferred knowledge can be extracted from the data leading to a more focused, efficient design procedure through enhanced process understanding. If this were to also lead to a reduction in the work involved in the design process, significant financial benefits would result. Furthermore by reducing time to market, manufacturing time under patent protection would increase.

Within this paper, the hypothesis considered is whether through the use of data mining techniques, it is possible to identify proteins that exhibit 'similar' characteristics to the current development organism/product based on the interrogation of protein sequence data. It is subsequently conjectured that through the use of information available on these past products and the manufacturing route selected, the development time of the new process could be reduced. Furthermore potential bottle-necks and challenges in production could be identified before they occur and hence areas of research that necessitate specific attention in the design process identified.

The approach presented is based on the interrogation of protein sequence data through the application of the clustering and classification algorithm, the Self-Organising Map (SOM) (Kohonen, 1982). The objective was to compare proteins using molecular level descriptors generated from the primary sequence as it was conjectured that it is the relationships at the molecular level that influence the tertiary structure of the protein and ultimately which then define the process and hence how it will behave at the production scale. Using this information, i.e. which proteins are similar, it is then possible to look at the properties of these proteins.

Specifically, this paper focuses on the causes of variation in downstream processing (DSP) requirements. Two case studies are considered. Case Study 1 is concerned with identifying whether differences between the likely downstream processing steps required for primary purification could be determined from the primary sequence. Case Study 2 relates to inclusion body formation. Specifically whether it is possible to pre-determine the appropriate solvent to use to re-dissolve proteins, which have formed inclusion bodies during the fermentation process.

2. Methodology

The data utilised was provided by Avecia Biologics, and comprised of amino acid sequences from 41 of their previously developed proteins (Table 1). These sequences exhibit a great deal of variation, particularly in terms of their size (6.2–92.7 KDa), pI (4.34–10.09) and hydrophobicity (GRAVY range of −1.044–0.123). In addition to the sequence data, information relating to the DSP steps was also provided. In Case Study 1, information pertaining to whether an expanded-bed was used or more traditional separation units such as filtration and centrifugation was the basis for classification of the proteins. In Case Study 2, information related to whether proteins had experienced inclusion body formation in DSP. This DSP data was used to generate class descriptors for each protein.

Using the sequence data, a number of protein descriptors were generated as described in the research reported by Idicula-Thomas et al. (2006). The primary sequence of each of Avecia's proteins was presented to the ProtParam tool on the ExPASy World Wide Web server (Gasteiger et al., 2005). This on-line analysis tool utilises a set of empirical relationships to calculate a number of protein parameters directly from the primary sequence. Parameters calculated include: Theoretical Isoelectric Point (pI), Instability Index (II) (Guruprasad et al., 1990), Aliphatic Index (Ikai, 1980) and Grand Average of Hydropathy (GRAVY) (Kyte and Doolittle, 1982). Alongside this information, the on-line tool also provides a breakdown of the molecular weight and associated amino acid abundances for the protein, which were used to generate di-peptide and tri-peptide scores for each protein. This data was assembled into a database of protein characteristics that could be used as descriptors for the classification of the proteins with regards to the associated class data. A total of 12,732 variables were attained (Table 2).

Most classification algorithms are unable to handle such a large number of variables effectively, due to issues such as computational load. To address this challenge, methods for reducing the number of variables can be applied. Idicula-Thomas et al. (2006) utilised the unbalanced correlation score (Weston et al., 2003b) to identify the top 20 features correlated with solubility. In this research a number of feature selection algorithms were investigated including Fisher correlation score, unbalanced correlation score (Weston et al., 2003b), recursive feature elimination (Guyon et al., 2002) and primal zero-norm (concave minimisation) (Bradley and Mangasarian, 1998). The detailed results of the study are not reported but it was observed that some feature selection algorithms lead

Table 1 – Primary sequence identifiers for proteins used in Case Study 1 & 2.

Protein Number	Case Study 1 sequence IDs	Case Study 2 sequence IDs	Protein description
1	B2007		Protease inhibitor
2	B2033		Enzyme
3	B2109		Protease inhibitor
4	B2114 1a	B2114 1a	Growth factor binding protein
5	B2114 2	B2114 2	Growth factor
6	B2187		Trefoil protein
7	B2212		Anti viral protein
8	B2257	B2257	Cytokine
9	B2272	B2272	Antigen
10	B2285		Metallopeptidase
11	B2289	B2289	Enzyme
12	B2296		Chemo-attractant protein
13	B2337		GST protease fusion protein
14	B2346		Heat shock antigen fusion protein
15	B2359		Antibody based fusion protein
16	B2365/F		Antigen
17	B2365/V		Antigen
18	B2377		Enterotoxin
19	B2385	B2385	Adipocyte protein fragment
20	B2403		Antibody fragment
21	B2407		Fibronectin derived protein
22	B2422		Metalloprotease
23	B2428	B2428	Cytokine
24	B2436		Enzyme
25	B2438		Growth factor
26	B2454		Cytokine
27	B2462	B2462	Growth factor/Cytokine
28	B2463		Enzyme fusion protein
29	B2484	B2484	Antiviral protein
30	B2494		Flagellin antigen fusion protein
31	B2521		Cytokine
32	B2530	B2530	Endopeptidase
33	B2531	B2531	Growth factor/Cytokine
34	B2547	B2547	Antigen
35	B2550		Cytokine
36	B2569		Enzyme
37	B2588		Growth factor HSA fusion protein
38	B2610	B2610	Growth factor/Cytokine
39	B7013		Protease inhibitor
40	B8003		Complement inhibitor
41	B9043		Transferrin family protein

to variables incorrectly being identified as significant due, in part, to chance correlations occurring as a consequence of the large number of variables in the initial data set. The Fisher correlation score (Weston et al., 2003a) was utilised as it gave the most reliable results following discussions with the experts at Avecia Biologics. The Fisher correlation score calculates a rank of the relative importance of each variable independent of the other variables:

$$f_j = \frac{(\mu_{j_{(+)}} - \mu_{j_{(-)}})^2}{(\sigma_{j_{(+)}})^2 + (\sigma_{j_{(-)}})^2} \quad (1)$$

where $\mu_{j(+)}$ and $\mu_{j(-)}$, are the mean values of variable j for the positively and negatively classified samples respectively; $\sigma_{j(+)}$ and $\sigma_{j(-)}$ are the corresponding standard deviations.

Table 2 – Variables contained in the initial data set generated from the primary sequences by the ExPASy World Wide Web server's ProtParam tool.

Variable number	Parameter details
1	Number of amino acids
2	Molecular weight
3	Theoretical pI
4	Net charge
5	Number of carbons atoms
6	Number of hydrogen atoms
7	Number of nitrogen atoms
8	Number of sulphur atoms
9	Extinction coefficient A280 (all half Cys)
10	Instability index
11	Aliphatic index
12	Grand average of hydropathy (GRAVY)
13–35	Amino acid abundances (23 inc. non-standard AA's: B, X and Z)
36–564	Amino acid di-peptide abundances (23^2 combinations)
565–12,732	Amino acid tri-peptide abundances (23^3 combinations)

Once the top features ranked by Fisher correlation score had been identified, the self-organising map (Kohonen, 1982) unsupervised learning technique was used to cluster the proteins. An unsupervised approach aims to learn how to represent particular input patterns in a way that reflects the underlying structure of the overall collection of input patterns. In contrast to a supervised approach, there are no explicit target outputs associated with each input and hence the algorithm is not influenced by a priori information. A non-linear approach was adopted as the assumption of linearity is without basis when considering the complexity of a biological system. The SOM algorithm can handle both linear and non-linear systems and is thus preferable to a strategy that is solely applicable to linear systems.

The objective of the SOM is to map high dimensional input data onto a two-dimensional arrangement of nodes known as the feature space. In this feature space, each node is associated with a parametric vector in the real space, known as a codebook vector. Each of these vectors have the same order of dimensionality as the training samples. The resulting feature space projection is then utilised for classification purposes by overlaying the training data on the map and assigning each node an associated class based on the relative abundance of the samples for which that node is the best matching unit (BMU). The BMU is determined by calculating the minimum Euclidean distance from each codebook vector to each training sample. Since the SOM algorithm provides a 'spatial ordering' in the feature space, proteins with similar properties will typically lie close to each other on the map. The feature space map can then be used to give an indication of the likely classification of unseen data by calculating the BMU for the new protein and assigning the BMU's class to the new protein.

3. Results and discussion

The results presented in this section were produced using the MATLAB SOM toolbox (Vesanto et al., 2000). Two case studies are presented, which focus on influencing decisions related to the downstream processing of therapeutic proteins.

Table 3 – Top 5 features relative to the primary capture step ranked by Fisher correlation score.

Feature rank	Fisher score
1	Aliphatic index
2	Lysine–glutamic acid–cysteine (KEC)
3	Glutamine–threonine (QT)
4	Instability index
5	Glutamine–proline (QP)

3.1. *Case Study 1*

Case Study 1 is concerned with identifying whether differences between contrasting downstream processing steps required for primary purification could be determined from the interrogation of the primary sequence data. The problem is related to the primary capture steps of the 41 previous processes developed at Avecia Biologics. Six of these projects utilised an Expanded Bed Adsorption (EBA) chromatography step rather than the more traditional approach of solids removal by centrifugation or filtration followed by Ion Exchange (IEX) chromatography or Hydrophobic Interaction Chromatography (HIC) for product capture and purification. The objective was to examine whether these six projects exhibited any degree of clustering, and if observed, was it possible to deduce which factors were predictors of this behaviour so this information could be used when developing future processes.

3.1.1. *Results and discussion*

The first stage was to apply the Fisher correlation score and identify the top 5 features, Table 3. A 4-fold cross validation of the 41 samples was then performed by classifying the data into two groups; those samples where the EBA was used and those that utilised an alternative strategy. The ratio of EBA processes to non-EBA processes, in each cross-validation subset was kept approximately constant to ensure a fair comparison. The resulting cross-validation produced an average percentage of correct classification of 72.5% with the individual cross validations giving results of 66.7%, 80.0%, 70.0% and 73.3%. It should be noted that validation samples assigned to undefined regions of the map (empty nodes) were not included in the correct classification category.

Fig. 1 shows a typical feature space projection resulting from one of the cross validation runs. The size of the feature space map is computed by utilising the heuristic formula:

$$N = 5 * \sqrt{n} \quad (2)$$

where N is the number of nodes and n is the number of training samples. The length and width of the map are then determined by calculating the square root of the ratio of the two largest eigenvalues of the covariance matrix of the training samples (Vesanto et al., 2000).

On the left side of Fig. 1 is a feature space plot coloured in terms of the primary capture steps of the training data. Dark grey nodes describe regions of the map which are dominated by training samples which utilised EBA as the primary capture mechanism, whereas light grey nodes represent regions which describe an alternative primary capture step. Black nodes represent regions of the map where no training data exists. As can be seen, there is a distinctive clustering of those processes that utilise the EBA step. This is captured by the dark grey cluster in the upper section of the map and is further supported by the cluster of empty nodes directly surrounding this cluster. On the right side of Fig. 1, the size of the pie charts represents the total number of protein samples in that region of the map, with the sections of the pie representing the abundance of each individual class. From this figure it is evident that in general the membership of these codebook vectors is uniquely associated with a specific process with only 2 codebook vectors containing samples associated with both classes.

The Unified Distance matrix (U-matrix), Fig. 2 visualises the distance between neighbouring nodes on the feature space plot and hence has one additional node in each direction of the

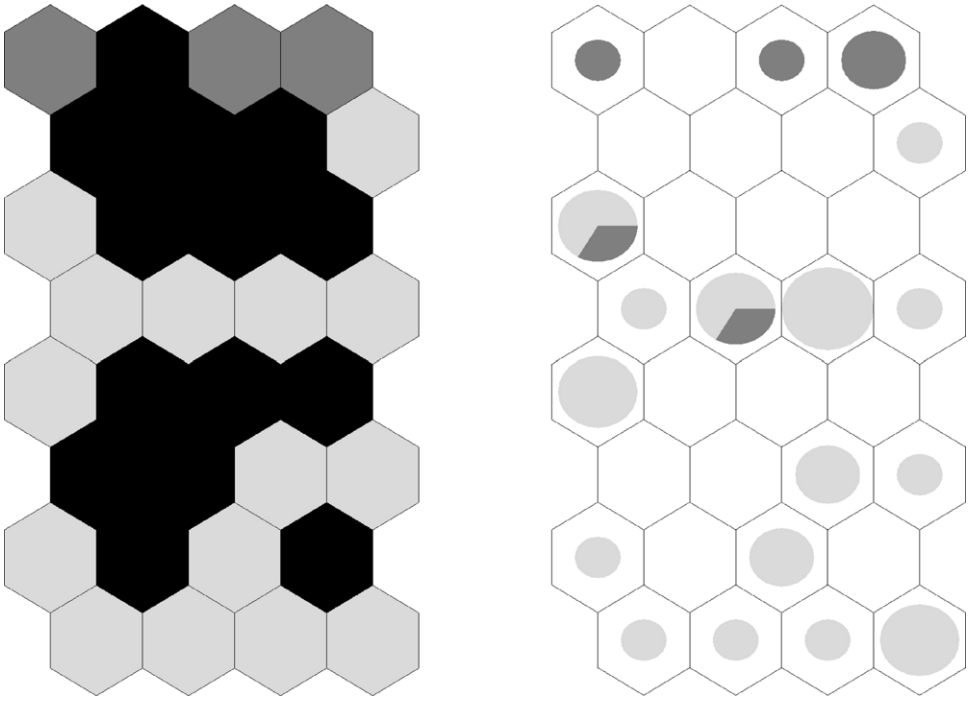

Fig. 1 – Typical feature space projection for Case Study 1 with associated sample distribution of nodes (Dark grey node – EBA; Light grey – alternative method; Black node – Empty).

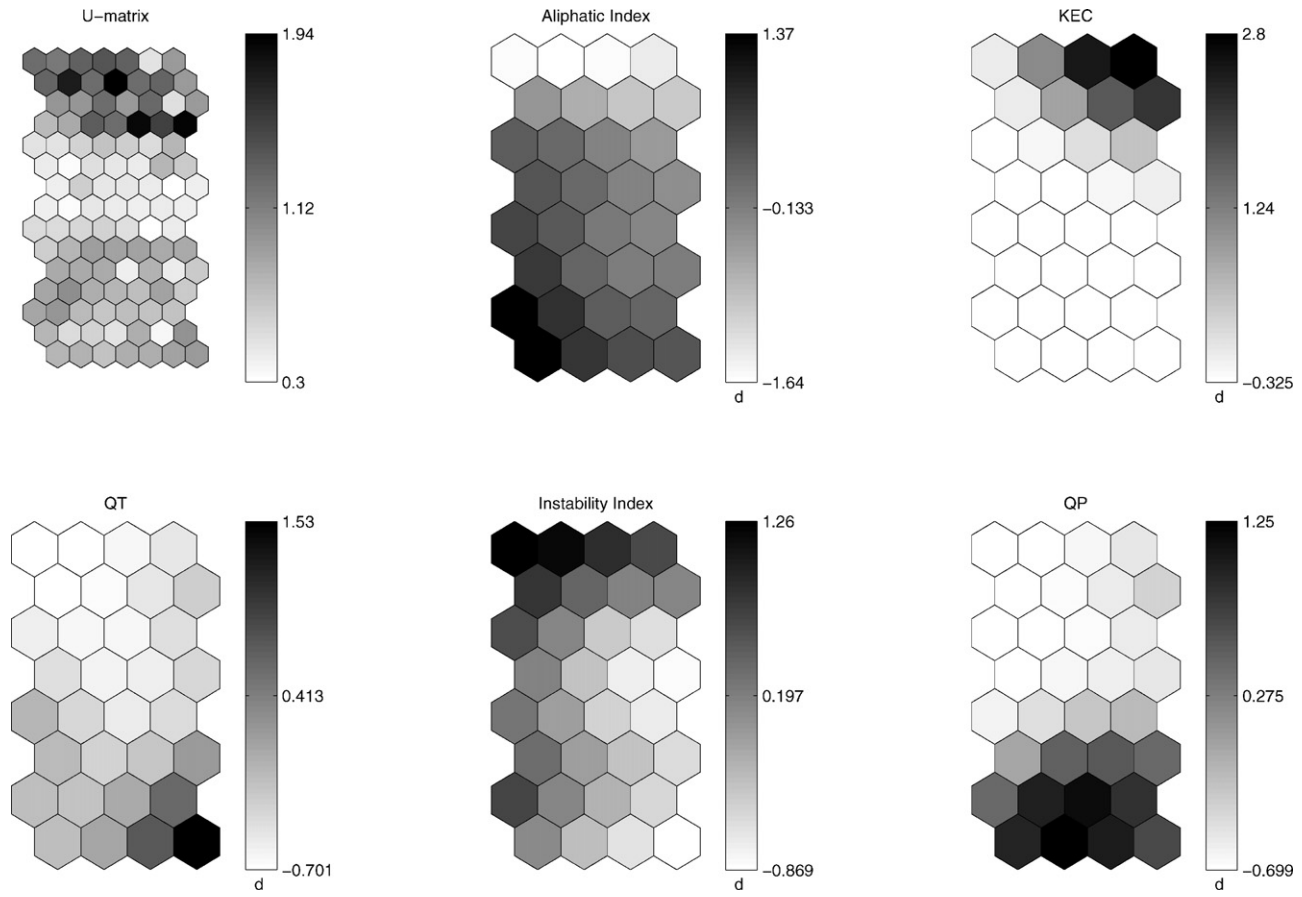

Fig. 2 – Unified distance matrix and individual variable contributions associated with Fig. 1.

map. The individual contribution plots show how each variable relates to the feature space plot. The grey scale gradient bar represents the range of values observed for each variable from high to low (black to white). Analysis of the individual contribution plots associated with the feature space map, Fig. 2, provides a more detailed explanation of the relationship between the top 5 features and the output classification. The heat map of Aliphatic Index, for example, shows a strong inverse correlation with the primary recovery technique, i.e. proteins with a small value for their aliphatic index are more likely to utilise an EBA step in the recovery process. Conversely, the heat map for Instability Index shows a strong positive correlation with the primary recovery technique, which means proteins which have an EBA step in the down-stream processing stage tend to have a larger value for the instability index than those processed by more traditional routes.

These results demonstrate that the combination of a feature selection technique and the SOM algorithm to identify clusters and enable visualisation of high dimensional data can provide useful information with regard to design protocols for future processes.

3.2. *Case Study 2*

Case Study 2 was concerned with inclusion body formation. Specifically whether it was possible to pre-determine the appropriate solvent to use to re-dissolve proteins, which have formed inclusion bodies during the fermentation process, into solution. The problem considers a subset of 13 proteins from previous processes where inclusion bodies were formed during the fermentation process. For the majority of these projects Urea has been sufficient for the solubilisation of inclusion bodies, whilst in a small number of projects the stronger chaotrophic agent Guanidine was required. It is known that these molecules disrupt the hydrophobic interactions, but the manner in which they do is not well understood (Voet and Voet, 1995). Avecia Biologics were interested in what features of the proteins may account for these differences.

3.2.1. *Results and discussion*

As in the previous case study, the top 5 features ranked by Fisher correlation score were selected, Table 4. A 4-fold cross validation of the 13 samples was performed by classifying the data into two groups; those samples that utilised Urea and those that utilised Guanidine. The resulting cross-validation produced an average percentage of correct classification of 83.3%, with individual results being 100%, 66.7%, 100% and 66.7%.

Fig. 3 shows a typical feature space projection resulting from the one of the cross validation runs. On the left side of the figure is the feature space plot coloured in terms of the solubilisation agent used. Dark grey nodes describe regions of the map that are dominated by training samples which utilise Guanidine, whereas light grey nodes represent regions

Table 4 – Top 5 features relative to the inclusion body formation ranked by Fisher correlation score.

Feature rank	Fisher score
1	Threonine–histidine (TH)
2	Glycine–glutamine (GQ)
3	Net charge
4	Proline–alanine (PA)
5	Glutamic acid–arginine (ER)

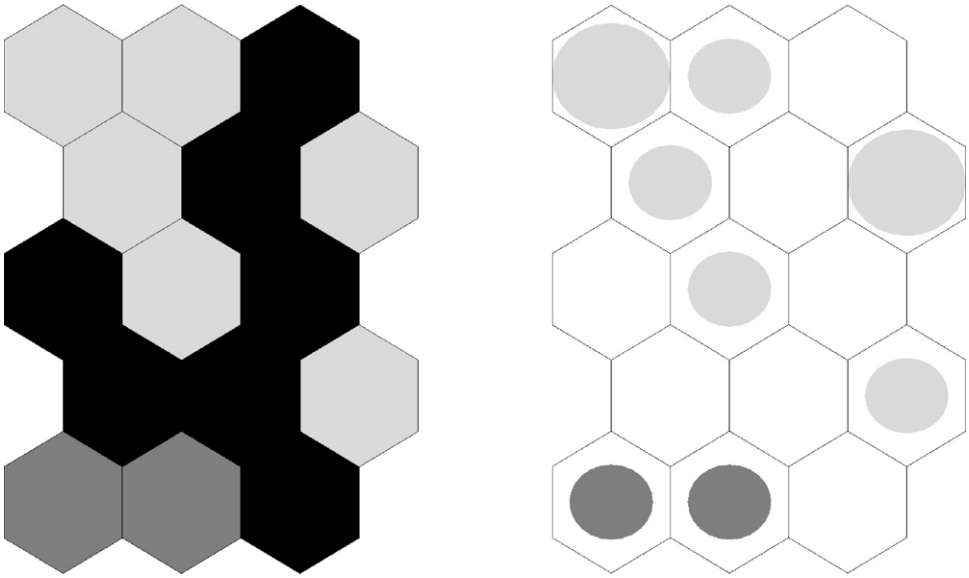

Fig. 3 – Typical feature space projection for Case Study 2 with associated sample distribution of nodes (dark grey nodes – urea based samples; light grey nodes – guamidine; black – no training data present).

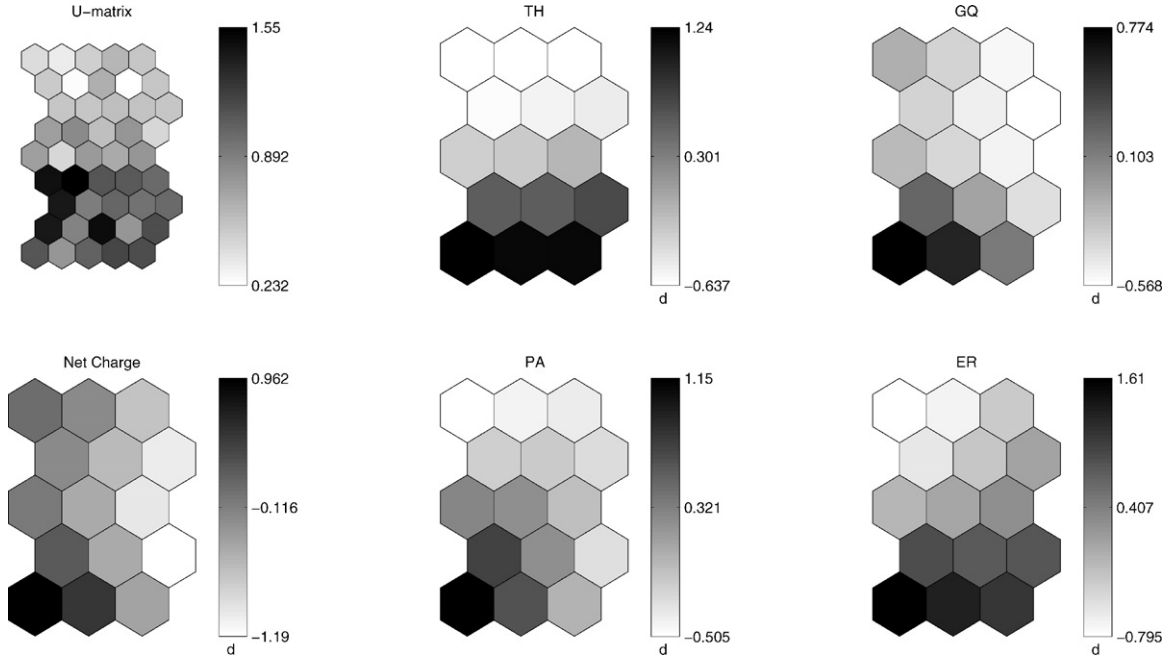

Fig. 4 – Unified distance matrix and individual variable contributions associated with Fig. 3.

where samples utilised Urea. From the figure on the right, it can be observed that no regions of mixed classification exist, and hence the pie charts are whole. Fig. 4 shows the U-matrix and contribution plots for each variable used in the training of the map. From the feature space map in Fig. 4, samples which group towards the lower left corner of the map typically exhibit high values for Net Charge. This result provides evidence to suggest that the net charge of a protein has an influence on what conditions are required for solubilisation. Specifically, proteins with a strong negative net charge are more likely to utilise guanidine for solubilisation. This finding is reinforced by the appearance of the Glutamic Acid–Arginine (ER) dipeptide, as these polar amino acids both contribute to the overall net charge of the protein. Further investigation led to the hypothesis that it may be that charged chaotrophic agents, such as the guanidinium ion found in guanidine, are required to overcome ionic attraction effects in highly charged proteins. This could explain why in situations where a protein is insoluble in the non-ionic agent urea, they are likely to be soluble in the stronger chaotrophic agent, guanidine.

4. Conclusions

The case studies presented in this paper provide clear evidence that information extracted from protein sequence data can be exploited to aid process route selection. By utilising the visualisation properties of the SOM algorithm, in conjunction with feature selection techniques, it has been possible to cluster proteins in terms of their similarity with respect to different classification criteria. This similarity information, coupled with the expert knowledge of process design engineers, can be used to aid decision making during development of new manufacturing scale processes. The average classification performance on cross-validation for the two case studies

was 72.5% and 83.3%. These results were considered good due to the small sample size. The need to apply feature selection prior to clustering to ensure spurious correlations do not materialise is an essential stage in the analysis.

Finally, while the concept of inferring knowledge across projects is demonstrated here, it is important to re-enforce the need for highly skilled individuals to interpret the information obtained appropriately and to utilise it effectively in the design process.

Acknowledgements

The authors wish to thank Avecia Biologics for providing access to their protein sequence data, along with the technical expertise and additional funding which enabled this research. This research was supported by EPSRC grant number GR/T11364/01 and an Engineering Doctorate from the UCL EngD for Bioprocess Leadership.

References

Avramenko, Y., Kroslawski, A., 2006. Similarity concept for case-based design in process engineering. Computers and Chemical Engineering 30, 548–557.

Bradley, P.S., Mangasarian, O.L., 1998. Feature selection via concave minimization and support vector machines. In: Fifteenth International Conference on Machine Learning, pp. 82–90.

Gasteiger, E., Hoogland, C., Gattiker, A., et al., 2005. Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (Ed.), The Proteomics Protocols Handbook. Humana Press.

Guruprasad, K., Reddy, B.V.B., Pandit, M.W., 1990. Correlation between stability of a orotein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Engineering 4, 155–161.

Guyon, I., Weston, J., Barnhill, S., et al., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46 (1), 389–422.

Idicula-Thomas, S., Kulkarni Kulkarni, A.J., et al., 2006. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. Bioinformatics 22 (3), 278–284.

Ikai, A.J., 1980. Thermostability and aliphatic index of globular proteins. Journal of Biochemistry 88, 1895–1898.

Kohonen, T., 1982. Self-organizing formation of topologically correct feature maps. Biological Cybernetics 43, 59–69.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157, 105–132.

Micheletti, M., Lye, G.J., 2006. Microscale bioprocess optimisation. Current Opinion in Biotechnology 17 (6), 611–618.

Shukla, A.A., Hubbard, B., Tressel, T., et al., 2007. Downstream processing of monoclonal antibodies – application of platform approaches. Journal of Chromatography B 848 (1), 28–39.

Vesanto, J., Himberg, J., Alhoniemi, E., et al., 2000. SOM Toolbox for MATLAB 5. Helsinki University of Technology.

Voet, D., Voet, J., 1995. Biochemistry. John Wiley & Sons, England.

Walsh, G., 2006. Biopharmaceutical benchmarks 2006. Nat. Biotechnol. 24 (7), 769–776.

Weston, J., Elisseeff, A., Scholkopf, B., et al., 2003a. Spider Toolbox for MATLAB.

Weston, J., Perez-Cruz, F., Bousquet, O., et al., 2003b. Feature selection and transduction for prediction of molecular bioactivity for drug design. Bioinformatics 19 (6), 764–771.